ガウス分布による単語と句の意味の分布的表現

島岡 聖世‡　村岡 雅康†　山本 風人†　渡邉 陽太郎†　岡崎 直観†*　乾 健太郎†
東北大学†‡　科学技術振興機構さきがけ*
simaokasonse@yahoo.co.jp‡
{muraoka, kazeto, yotaro-w, okazaki, inui}@.ecei.tohoku.ac.jp†

1 はじめに

分布意味論 (Distributional Semantics) と呼ばれる考え方によれば、テキスト集合から得られた各単語に対する共起頻度の統計において、よく似た分布を持つ単語は互いによく似た意味を持つ。このような考え方に基く、大規模テキストデータから得られた単語の共起情報を用いて単語の意味表現を実数ベクトルとして表す試みは、Web の発達等により大量のテキストが容易に利用できるようになった近年、活発に議論されてきた。このような手法によって学習された意味表現は**単語ベクトル**などと呼ばれ、単語の統語的・意味的な特性を内包するとともに、実数で表現されているために計算機上での扱いが容易である点や、大量のテキストさえあれば自動的に構築が可能な点など、数多くの利点を備えている。

しかし、単語の意味をこのようなベクトル空間上の一点として表す手法には、原理上の限界がいくつか考えられる。まず第一に、単語ベクトルでは単語が持つ意味の広がりを捉えることが出来ない。例えば、単語 “good” と単語 “tasty” については、部分的に共通した意味を持ちつつも、“good” の方がより抽象的で広い意味を持つと考えられる。このように、各単語が示す概念にはある種の広がりが存在しており、その規模は単語によってそれぞれ異なる。しかし、単語ベクトルの枠組みでは単語の意味はベクトル空間上のある一点として表現されるために、このような関係を捉えることが出来ない。

第二に、単語ベクトルは文脈が与えられたときの単語の意味の変化を表すことが困難である。例えば、“catch a ball”、“catch a disease” という 2 つの表現を考える。このとき、前者の “catch” は意味的には “grab” に近いが、後者の “catch” は “contract” という単語の意味に近くなる。このように、単語の意味やニュアンスは出現する文脈に応じて変化する。意味が変化すれば、ベクトル空間上での位置も変化すると考えるのが自然であるが、単語ベクトルはこの変化がどのようなものであるかに関して何も伝えない。単語ベクトルに関する応用として、句や文を構成する持つ意味をそれらが含む単語のベクトルから構成的に計算するという試みがあるが [5, 6, 3]、このような応用において上で述べた限界は重大な問題であると考えられる。

そこで本稿では、これらの限界に対処するために、単語や句の意味表現を得る過程を、一貫して確率的演算として扱うモデルを提案する。このモデルでは、単語の持つ意味はベクトルではなく、図 1 のようにガウス分布により表現され、分布の平均は単語の典型的な意味を、分布の共分散は意味の広がりを表す。さらに、文脈が与えられたときの単語の意味も同様にしてガウス分布で表す。

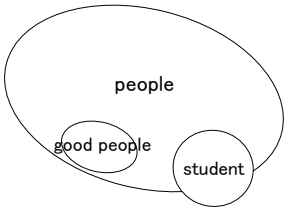


図 1: ガウス分布による意味表現

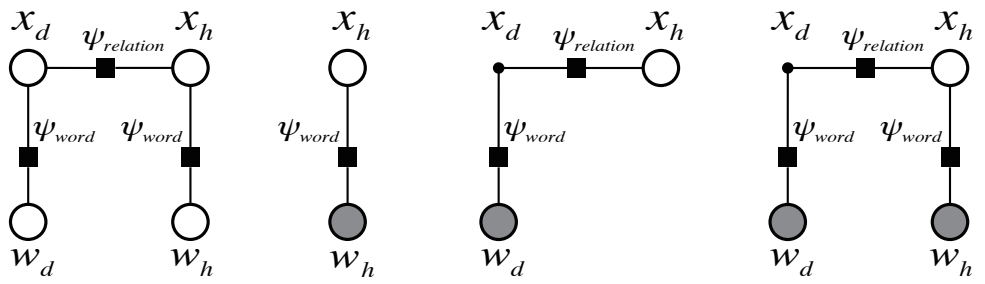
2 提案手法

本節では提案手法の概要を述べる。提案手法は、任意の単語数の句や文に対して適用できる一般的なものであるが、本稿では簡単のために 2 単語からなる句に限定して議論を行う。主要語 w_h と従属語 w_d が、依存構造上で句を構成しているとする。このとき、句におけるそれぞれの単語の意味が、D 次元実数ベクトル $\boldsymbol{x}_h$、$\boldsymbol{x}_d$ で表されるとする。$\boldsymbol{x}_h$、$\boldsymbol{x}_d$ を確率的に推定することは、確率分布 $P(\boldsymbol{x}_h, \boldsymbol{x}_d | w_h, w_d, \boldsymbol{\Theta})$ を計算することに対応する。$\boldsymbol{\Theta}$ はパラメータの集合である。$\boldsymbol{\Theta}$ を最適化するためには、$\boldsymbol{x}_h$、$\boldsymbol{x}_d$ に該当する観測データが必要であるが、これは存在しない。そこで本稿では、間接的な最適化方法を採用する。すなわち、単語の意味はその用法によって定まるとする分布的意味論の仮説に基づき、単語 w_h、w_d を確率変数として捉えた同時分布 $P(\boldsymbol{x}_h, \boldsymbol{x}_d, w_h, w_d | \boldsymbol{\Theta})$ を考え、$P(\boldsymbol{x}_h, \boldsymbol{x}_d | w_h, w_d, \boldsymbol{\Theta})$ の代わりに同時分布から $\boldsymbol{x}_h$ と $\boldsymbol{x}_d$ を積分消去した確率分布 $P(w_h, w_d | \boldsymbol{\Theta})$ を最適化する。この確率分布は言語モデルと呼ばれる。言語モデルのパラメータは、実際のテキストに現れる句を訓練データとして用いて学習することができる。$\boldsymbol{x}_h$ と $\boldsymbol{x}_d$ は言語モデルの隠れ変数とみなすことができる。

2.1 マルコフ確率場によるモデル

具体的なモデルをマルコフ確率場によって与える。同時分布 $P(\boldsymbol{x}_h, \boldsymbol{x}_d, w_h, w_d | \boldsymbol{\Theta})$ の確率場は図 2(a) のように表される。図 2 はグラフィカルモデルにおいて、因子グラフと呼ばれる記法に従っている。丸いノードは確率変数を表し、単語ノード w_h、w_d と意味ノード $\boldsymbol{x}_h$、$\boldsymbol{x}_d$ からなる。黒い四角のノードは因子関数であり、2 つの確率変数の間の相互作用を規定する。因子関数には 2 つの種類があり、それぞれ ψ_{word}、$\psi_{relation}$ と表す。

提案モデルにおいて、文脈が与えられていない状態での主要語 w_h の意味を表す確率場は図 2(b) で表される。灰色のノードは観測変数を表す。また、従属語 w_d の主要語 w_h に対する修飾作用を表す確率場は図 2(c) で表される。黒く小さいノードは変数が積分消去されたこと



(a) 同時分布　(b) 単独での単語の意味　(c) 従属語による修飾作用　(d) 文脈における単語の意味

図 2: 因子グラフによる提案モデル

を表す。主要語と従属語の相互作用により、文脈における主要語の意味が定まる過程を表すのが図 2(d) である。

ψ_{word} は、単語とその意味の相互作用を規定する因子関数であり、単語 w に固有の 2 つのパラメータ $\boldsymbol{\mu}_w$、$\boldsymbol{\Lambda}_w$ を用いて次のように定義される：

$$\psi_{word}(\boldsymbol{x}, w) = \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_w)^T\boldsymbol{\Lambda}_w(\boldsymbol{x}-\boldsymbol{\mu}_w)\right) \quad (1)$$

$\boldsymbol{\mu}_w$ は単語 w の典型的な意味を表す D 次元の実数ベクトルである。これは、既存の研究における、単語ベクトルに相当する。また、非特異な $D \times D$ 次元正定値対称行列 $\boldsymbol{\Lambda}_w^{-1}$ は単語の意味の広がりを表すパラメータである。$\boldsymbol{\mu}_w$、$\boldsymbol{\Lambda}_w^{-1}$ をそれぞれ**平均パラメータ**、**共分散パラメータ**と呼ぶ。因子関数 ψ_{word} の役割は、単語の文脈における意味が、単語の典型的な意味に対して、どのように変化しうるかを制御することである。

一方、$\psi_{relation}$ は、w_h の意味 $\boldsymbol{x}_h$ と w_d の意味 $\boldsymbol{x}_d$ の間の相互作用を規定する。例えば 2 単語間の修飾関係を r とすると、$\psi_{relation}$ は r に固有のパラメータ $\boldsymbol{W}_r$ を用いて次のように定義される：

$$\psi_{relation}(\boldsymbol{x}_h, \boldsymbol{x}_d) = \exp\left(-\frac{1}{2}(\boldsymbol{x}_h - \boldsymbol{W}_r\boldsymbol{x}_d)^T(\boldsymbol{x}_h - \boldsymbol{W}_r\boldsymbol{x}_d)\right) \quad (2)$$

$D \times D$ 次元行列 $\boldsymbol{W}_r$ は、従属語 w_d の意味 $\boldsymbol{x}_d$ が主要語 w_h の意味 $\boldsymbol{x}_h$ に対して与える修飾作用を定める。積 $\boldsymbol{W}_r\boldsymbol{x}_d$ は、従属語が主要語に対して「期待する」意味を表していると考えることが出来る。

因子関数と変数について定義したので、確率分布を具体的な式で表すことが出来る。3 つの因子関数全ての積を、$\Psi_{all} = \Psi_{all}(\boldsymbol{x}_h, \boldsymbol{x}_d, w_h, w_d)$ とすれば、条件付き分布 $P(\boldsymbol{x}_h, \boldsymbol{x}_d | w_h, w_d, \boldsymbol{\Theta})$ は、

$$\begin{aligned} P(\boldsymbol{x}_h, \boldsymbol{x}_d | w_h, w_d, \boldsymbol{\Theta}) &= \frac{\Psi_{all}(\boldsymbol{x}_h, \boldsymbol{x}_d, w_h, w_d)}{\int \Psi_{all}(\boldsymbol{x}_h, \boldsymbol{x}_d, w_h, w_d)\ d\boldsymbol{x}_h d\boldsymbol{x}_d} \\ &= N\left(\begin{pmatrix}\boldsymbol{x}_h \\ \boldsymbol{x}_d\end{pmatrix} | \begin{pmatrix}\boldsymbol{m}_h \\ \boldsymbol{m}_d\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_{hh} & \boldsymbol{\Sigma}_{hd} \\ \boldsymbol{\Sigma}_{dh} & \boldsymbol{\Sigma}_{dd}\end{pmatrix}\right) \end{aligned} \quad (3)$$

というガウス分布となる。ただし、

$$\begin{pmatrix}\boldsymbol{\Sigma}_{hh} & \boldsymbol{\Sigma}_{hd} \\ \boldsymbol{\Sigma}_{dh} & \boldsymbol{\Sigma}_{dd}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Lambda}_{w_h} + \boldsymbol{I} & -\boldsymbol{W}_r \\ -\boldsymbol{W}_r^T & \boldsymbol{\Lambda}_{w_d} + \boldsymbol{W}_r^T\boldsymbol{W}_r\end{pmatrix}^{-1}$$

$$\begin{aligned} \boldsymbol{m}_h &= \boldsymbol{\Sigma}_{hh}(\boldsymbol{W}_r(\boldsymbol{\Lambda}_{w_d} + \boldsymbol{W}_r^T\boldsymbol{W}_r)^{-1}\boldsymbol{\Lambda}_{w_d}\boldsymbol{\mu}_{w_d} + \boldsymbol{\Lambda}_{w_h}\boldsymbol{\mu}_{w_h}) \\ \boldsymbol{m}_d &= \boldsymbol{\Sigma}_{dd}(\boldsymbol{W}_r^T(\boldsymbol{\Lambda}_{w_h} + \boldsymbol{I})^{-1}\boldsymbol{\Lambda}_{w_h}\boldsymbol{\mu}_{w_h} + \boldsymbol{\Lambda}_{w_d}\boldsymbol{\mu}_{w_d}) \end{aligned}$$

である ($\boldsymbol{I}$ は単位行列)。

言語モデル $P(w_h, w_d | \boldsymbol{\Theta})$ は、

$$P(w_h, w_d | \boldsymbol{\Theta}) = \frac{\int \Psi_{all}(\boldsymbol{x}_h, \boldsymbol{x}_d, w_h, w_d)\ d\boldsymbol{x}_h d\boldsymbol{x}_d}{\sum_{w_h, w_d}\int \Psi_{all}(\boldsymbol{x}_h, \boldsymbol{x}_d, w_h, w_d)\ d\boldsymbol{x}_h d\boldsymbol{x}_d} \quad (4)$$

となるが、これは語彙サイズの 2 乗に比例する計算量を必要とするため、計算量が多くなりすぎてしまい、厳密に求めることは困難である。

2.2 パラメータ推定

パラメータ推定は、確率的勾配上昇法を用いて行う。この際、言語モデルについての勾配を導く必要がある。ただし、計算量の問題から、正確な勾配を求めることができないため、モンテカルロ法による近似を行う。

言語モデルは式 (4) で定義され、その対数の勾配は次のようになる：

$$\begin{aligned} &\frac{\partial \ln P(w_h, w_d | \boldsymbol{\Theta})}{\partial \boldsymbol{\theta}} \\ &= \int \left(\frac{\partial \ln \Psi_{all}}{\partial \boldsymbol{\theta}}\right) P(\boldsymbol{x}_h, \boldsymbol{x}_d | w_h, w_d, \boldsymbol{\Theta})\ d\boldsymbol{x}_h d\boldsymbol{x}_d \\ &- \underbrace{\sum_{w_h, w_d}\int \left(\frac{\partial \ln \Psi_{all}}{\partial \boldsymbol{\theta}}\right) P(\boldsymbol{x}_h, \boldsymbol{x}_d, w_h, w_d | \boldsymbol{\Theta})\ d\boldsymbol{x}_h d\boldsymbol{x}_d}_{\text{Ⓐ}} \end{aligned} \quad (5)$$

上式において Ⓐ といた部分は、同時分布を計算する必要があるが、これは言語モデルを上回る計算コストがかかるため、計算は難しい。そこで、次のようなモンテカルロ法による近似を行う：

$$\begin{aligned} \text{Ⓐ} &= \sum_{w_h, w_d}\int \left(\frac{\partial \ln \Psi_{all}}{\partial \boldsymbol{\theta}}\right) P(\boldsymbol{x}_h, \boldsymbol{x}_d, w_h, w_d | \boldsymbol{\Theta})\ d\boldsymbol{x}_h d\boldsymbol{x}_d \\ &\approx \int \left(\frac{\partial \ln \Psi_{all}}{\partial \boldsymbol{\theta}}\right) P(\boldsymbol{x}_h, \boldsymbol{x}_d | w_h', w_d', \boldsymbol{\Theta})\ d\boldsymbol{x}_h d\boldsymbol{x}_d \end{aligned} \quad (6)$$

となる。ただし、w_h'、w_d' は同時分布からのギブスサンプルである。結局、式 (5) は以下のように近似される。

$$\begin{aligned} &\frac{\partial \ln P(w_h, w_d | \boldsymbol{\Theta})}{\partial \boldsymbol{\theta}} \\ &\approx \int \left(\frac{\partial \ln \Psi_{all}(\boldsymbol{x}_h, \boldsymbol{x}_d, w_h, w_d)}{\partial \boldsymbol{\theta}}\right) P(\boldsymbol{x}_h, \boldsymbol{x}_d | w_h, w_d, \boldsymbol{\Theta})\ d\boldsymbol{x}_h d\boldsymbol{x}_d \\ &- \int \left(\frac{\partial \ln \Psi_{all}(\boldsymbol{x}_h, \boldsymbol{x}_d, w_h', w_d')}{\partial \boldsymbol{\theta}}\right) P(\boldsymbol{x}_h, \boldsymbol{x}_d | w_h', w_d', \boldsymbol{\Theta})\ d\boldsymbol{x}_h d\boldsymbol{x}_d \end{aligned}$$

この式は観測データにより条件付けされた分布の期待値から、ギブスサンプルにより条件付けされた分布の期待値を引いたものである。勾配をこのように近似する学習法は CD 学習と呼ばれている [4]。

3 アルゴリズム

データセット中の 1 つの句に対しての実際の確率的勾配上昇法の 1 ステップを以下に示す。

①句を構成する 2 つの単語を w_h、w_d とする。

②単語 w_h、w_d に対して、ガウス分布の統計量

$\boldsymbol{m}_h, \boldsymbol{m}_d, \boldsymbol{\Sigma}_{hh}, \boldsymbol{\Sigma}_{hd}, \boldsymbol{\Sigma}_{dd}$ を計算する。
③ギブスサンプル $w_h^{'}, w_d^{'}$ を得る (このとき、高速化のために 1 グラム確率を提案分布とするメトロポリス法によるサンプリングを行う [1])。サンプリングを表す式は以下のようになる。

$$\begin{aligned} \boldsymbol{x}_h^{'}, \boldsymbol{x}_d^{'} &\sim P(\boldsymbol{x}_h, \boldsymbol{x}_d | w_h, w_d, \boldsymbol{\Theta}) \\ w_h^{'}, w_d^{'} &\sim P(w_h, w_d | \boldsymbol{x}_h^{'}, \boldsymbol{x}_d^{'}, \boldsymbol{\Theta}) \end{aligned}$$

④ギブスサンプル $w_h^{'}$、$w_d^{'}$ に対して、②と同様にガウス分布の統計量 $\boldsymbol{m}_{h'}, \boldsymbol{m}_{d'}, \boldsymbol{\Sigma}_{h'h'}, \boldsymbol{\Sigma}_{h'd'}, \boldsymbol{\Sigma}_{d'd'}$ を計算する。
⑤各パラメータについて勾配の近似を以下のように計算する：

$$\begin{aligned}
\frac{\partial \ln P(w_h, w_d|\boldsymbol{\Theta})}{\partial \boldsymbol{W}_r} &\approx \boldsymbol{\Sigma}_{hd} + \boldsymbol{m}_h \boldsymbol{m}_d^T - \boldsymbol{W}_r(\boldsymbol{\Sigma}_{dd} + \boldsymbol{m}_d \boldsymbol{m}_d^T) \\
&- \boldsymbol{\Sigma}_{h'd'} - \boldsymbol{m}_{h'} \boldsymbol{m}_{d'}^T \\
&+ \boldsymbol{W}_r(\boldsymbol{\Sigma}_{d'd'} + \boldsymbol{m}_{d'} \boldsymbol{m}_{d'}^T) \\
\frac{\partial \ln P(w_h, w_d|\boldsymbol{\Theta})}{\partial \boldsymbol{\mu}_{w_h}} &\approx -\boldsymbol{\Lambda}_{w_h}(\boldsymbol{\mu}_{w_h} - \boldsymbol{m}_h) \\
\frac{\partial \ln P(w_h, w_d|\boldsymbol{\Theta})}{\partial \boldsymbol{\mu}_{w_d}} &\approx -\boldsymbol{\Lambda}_{w_d}(\boldsymbol{\mu}_{w_d} - \boldsymbol{m}_d) \\
\frac{\partial \ln P(w_h, w_d|\boldsymbol{\Theta})}{\partial \boldsymbol{\mu}_{w_h^{'}}} &\approx \boldsymbol{\Lambda}_{w_h^{'}}(\boldsymbol{\mu}_{w_h^{'}} - \boldsymbol{m}_{h'}) \\
\frac{\partial \ln P(w_h, w_d|\boldsymbol{\Theta})}{\partial \boldsymbol{\mu}_{w_d^{'}}} &\approx \boldsymbol{\Lambda}_{w_d^{'}}(\boldsymbol{\mu}_{w_d^{'}} - \boldsymbol{m}_{d'}) \\
\frac{\partial \ln P(w_h, w_d|\boldsymbol{\Theta})}{\partial \boldsymbol{\Lambda}_{w_h}} &\approx \frac{1}{2}(-\boldsymbol{\Sigma}_{hh} - \boldsymbol{m}_h \boldsymbol{m}_h^T + \boldsymbol{m}_h \boldsymbol{\mu}_{w_h}^T \\
&\quad + \boldsymbol{\mu}_{w_h} \boldsymbol{m}_h^T - \boldsymbol{\mu}_{w_h} \boldsymbol{\mu}_{w_h}^T) \\
\frac{\partial \ln P(w_h, w_d|\boldsymbol{\Theta})}{\partial \boldsymbol{\Lambda}_{w_d}} &\approx \frac{1}{2}(-\boldsymbol{\Sigma}_{dd} - \boldsymbol{m}_d \boldsymbol{m}_d^T + \boldsymbol{m}_d \boldsymbol{\mu}_{w_d}^T \\
&\quad + \boldsymbol{\mu}_{w_d} \boldsymbol{m}_d^T - \boldsymbol{\mu}_{w_d} \boldsymbol{\mu}_{w_d}^T) \\
\frac{\partial \ln P(w_h, w_d|\boldsymbol{\Theta})}{\partial \boldsymbol{\Lambda}_{w_h^{'}}} &\approx -\frac{1}{2}(-\boldsymbol{\Sigma}_{h'h'} - \boldsymbol{m}_{h'} \boldsymbol{m}_{h'}^T + \boldsymbol{m}_{h'} \boldsymbol{\mu}_{w_h^{'}}^T \\
&\quad + \boldsymbol{\mu}_{w_h^{'}} \boldsymbol{m}_{h'}^T - \boldsymbol{\mu}_{w_h^{'}} \boldsymbol{\mu}_{w_h^{'}}^T) \\
\frac{\partial \ln P(w_h, w_d|\boldsymbol{\Theta})}{\partial \boldsymbol{\Lambda}_{w_d^{'}}} &\approx -\frac{1}{2}(-\boldsymbol{\Sigma}_{d'd'} - \boldsymbol{m}_{d'} \boldsymbol{m}_{d'}^T + \boldsymbol{m}_{d'} \boldsymbol{\mu}_{w_d^{'}}^T \\
&\quad + \boldsymbol{\mu}_{w_d^{'}} \boldsymbol{m}_{d'}^T - \boldsymbol{\mu}_{w_d^{'}} \boldsymbol{\mu}_{w_d^{'}}^T)
\end{aligned}$$

⑥パラメータを勾配上昇法により更新する。ただし、意味の広がりを表すパラメータについては正定値という制約を保ちながら最適化するために、以下のような更新を行う (α を学習率とする)：

$$\begin{aligned}
\boldsymbol{\Lambda}_{w_h} &:= (\boldsymbol{I} + \alpha \frac{\partial \ln P(w_h, w_d|\boldsymbol{\Theta})}{\partial \boldsymbol{\Lambda}_{w_h}})^T \boldsymbol{\Lambda}_{w_h} (\boldsymbol{I} + \alpha \frac{\partial \ln P(w_h, w_d|\boldsymbol{\Theta})}{\partial \boldsymbol{\Lambda}_{w_h}}) \\
\boldsymbol{\Lambda}_{w_d} &:= (\boldsymbol{I} + \alpha \frac{\partial \ln P(w_h, w_d|\boldsymbol{\Theta})}{\partial \boldsymbol{\Lambda}_{w_d}})^T \boldsymbol{\Lambda}_{w_d} (\boldsymbol{I} + \alpha \frac{\partial \ln P(w_h, w_d|\boldsymbol{\Theta})}{\partial \boldsymbol{\Lambda}_{w_d}}) \\
\boldsymbol{\Lambda}_{w_h^{'}} &:= (\boldsymbol{I} + \alpha \frac{\partial \ln P(w_h, w_d|\boldsymbol{\Theta})}{\partial \boldsymbol{\Lambda}_{w_h^{'}}})^T \boldsymbol{\Lambda}_{w_h^{'}} (\boldsymbol{I} + \alpha \frac{\partial \ln P(w_h, w_d|\boldsymbol{\Theta})}{\partial \boldsymbol{\Lambda}_{w_h^{'}}}) \\
\boldsymbol{\Lambda}_{w_d^{'}} &:= (\boldsymbol{I} + \alpha \frac{\partial \ln P(w_h, w_d|\boldsymbol{\Theta})}{\partial \boldsymbol{\Lambda}_{w_d^{'}}})^T \boldsymbol{\Lambda}_{w_d^{'}} (\boldsymbol{I} + \alpha \frac{\partial \ln P(w_h, w_d|\boldsymbol{\Theta})}{\partial \boldsymbol{\Lambda}_{w_d^{'}}})
\end{aligned}$$

4 実験

本節では評価実験とその結果について述べる。

4.1 設定

前節で説明したアルゴリズムに基づいて、提案モデルの学習を行った。訓練に用いるデータセットは、テキストコーパス中に出現するバイグラムから、⟨ 名詞, 形容詞⟩、⟨ 名詞, 動詞⟩、⟨ 名詞, 名詞⟩ の 3 つのパターンのいずれかに当てはまり、かつ 50 回以上出現するようなものを抽出することで作成した。コーパスには ClueWeb09 を使用し、データセットとして約 40 万句からなる重複しないバイグラムの集合が得られた。モデルの学習は、データセット全体に対して確率的勾配上昇法の学習のステップを約 2000 回繰り返すことで行った。語彙数は 1 万、次元は $D = 25$ とした。パラメータの初期値は、分散パラメータ $\boldsymbol{\Lambda}_w^{-1}$ については単位行列とし、それ以外のパラメータについては微小な乱数とした。

4.2 評価

4.2.1 平均パラメータの学習

まず、単語を表すガウス分布の平均パラメータ $\boldsymbol{\mu}_w$ について、それぞれの単語の意味をうまく捉えられているかどうかを調査した。具体的には、学習後のモデルにおけるいくつかの単語に対して、語彙の中から分布の平均におけるコサイン類似度が最も高い 5 単語を取り出した。結果を表 1 に示す。似た意味を持つ単語同士が互いに近接していることから、平均パラメータについては単語の典型的な意味を一定の精度で学習できていることがわかる。

4.2.2 共分散パラメータの学習

提案手法により意味の広がりが学習されているかどうかを確認するため、頻出語上位 1000 単語のうち、共分散パラメータ $\boldsymbol{\Lambda}_w^{-1}$ の行列式が最も大きい 10 語と、最も小さい 10 語を取り出した。それぞれ結果を表 2、3 に示す。結果から、共分散パラメータは単語の出現頻度と大きな相関を持つことがわかる。この傾向は、提案手法において、共分散パラメータが意味の広がりを捉えるように学習されていることを示唆する。何故なら、意味の広い単語ほど様々な文脈で使用可能なために、出現頻度が高くなると考えられるからである。

4.2.3 上位-下位関係の学習

4.2.1 節及び 4.2.2 節の結果が示唆するのは、互いに似た意味を持つ単語は平均パラメータも近い値になることと、意味の広い単語ほど分散パラメータが大きくなることである。これらを総合すると、ガウス分布の包含関係 (一方のガウス分布がもう一方を覆うような関係) により単語の上位-下位関係を捉えることが出来る可能性が考えられる。そこで、可視化により、包含関係が上位-下位関係を捉えられているような例を探した。可視化は、まず人手で選んだ幾つかの単語の平均パラメータに対して主成分分析を行い、値が最大の 2 つの主成分に対応する固有ベクトルが張る平面を求め、その平面に単語のガウス分布による表現を楕円として射影することにより行った。楕円のスケールは見やすいように調節した。この可視化の結果の一部を図 3 に示す。結果から、単語が意味的カテゴリごとにクラスタを形成していると共に、“day” や “people” などの単語の表現が、それぞれ “sunday” や “student” などの単語の表現を包含していることがわかる。

表 1: コサイン類似度が高い上位 5 語

brother	march	cheese	china
mother	april	rice	indonesia
son	january	milk	russia
wife	may	butter	japan
father	june	juice	iran
doctor	november	fruit	pennsylvania

表 2: 共分散が最も大きい 10 単語

単語	行列式	頻出順位
system	7.14e+151	9
service	4.82e+108	7
work	8.20e+107	11
time	7.91e+101	8
company	4.11e+99	23
information	4.00e+97	10
be	2.38e+94	1
use	2.17e+94	3
way	7.09e+92	29
day	6.62e+92	32

表 3: 共分散が最も小さい 10 単語

単語	行列式	頻出順位
single	9.04e-122	368
impact	1.03e-112	749
significant	1.41e-111	921
download	1.52e-97	169
click	2.42e-97	787
deliver	3.82e-95	739
define	6.95e-92	736
associate	1.16e-90	996
cut	4.86e-85	629
status	6.34e-85	685

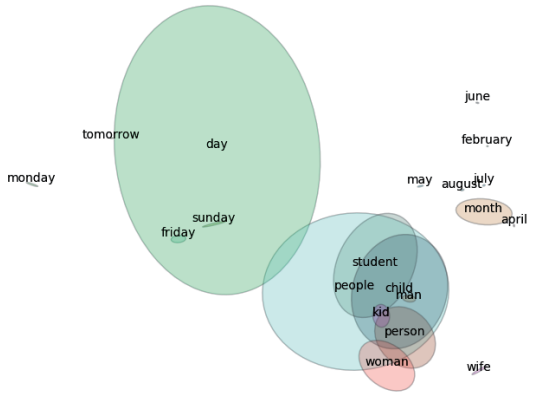


図 3: 主成分分析による可視化

4.2.4 文脈における意味の変化

次に、文脈が与えられたときに単語の意味が変化する様子を調べた。単語 “make” および “make” が述語である句の表現に対して、図 3 と同様の方法で可視化を行った。その結果を図 4 に示す。この図から、“make money” など “make” に目的語を与えられたときの句の共分散パラメータは “make” 単独での共分散パラメータにくらべて小さくなっていることがわかる。これは、提案モデルにより、文脈が与えられたときに単語の意味がより具体的になる過程が正しくモデル化されたことを表している。

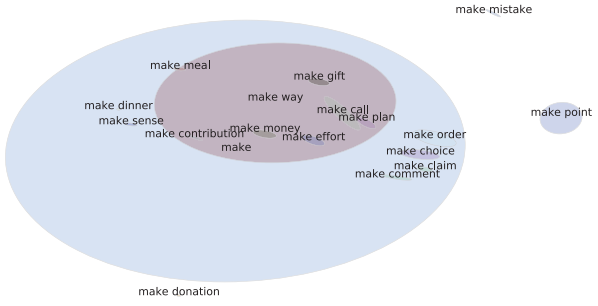


図 4: 主成分分析による可視化 2

5 先行研究

本節では、本研究に関連する先行研究を紹介する。分布意味論に基づいた単語の意味表現に関する研究は、単語の意味をベクトル座標として表すものが大半を占めている。本研究のように広がりを持った領域として単語の意味を表現するモデルは、我々の知る限りでは Erk らの提案したモデル [2] が唯一のものである。彼女らのモデルでは、我々のモデルと同様、単語の意味表現をベクトル空間上の領域によって表現し、領域の中心が単語の典型的な意味を、領域の広がりが単語の意味の変化の広さを表している。

我々の提案手法と Erk らの手法との違いとして、マルコフ確率場を用いて、単語だけでなく句や文の表現もガウス分布として表現していること、言語モデル学習により最適化を行っていることなどが挙げられる。

ところで、分布的意味論の枠組みにおける上位-下位関係を扱った概念として、分布的包含仮説 [7] がある。これは、単語 v の意味が単語 w の意味を含意するとき、v が出現するすべての典型的な文脈では、w もまた出現しうるという仮説である。これに関して、単語間の上位-下位関係と、単語を共起情報から学習されたガウス分布の包含関係の間には関連が存在すること、すなわち分布的包含仮説の正当性を示唆する結果が本研究の実験において得られたことは、興味深いことである。

6 結論

本稿では、単語と句の意味をガウス分布として表現するモデルを、パラメータ推定法と共に提案した。提案手法は、3 単語以上からなる句や文にも自然に拡張できるものである。また実験により、(i) 平均パラメータについては既存の単語ベクトルの表現と同様な振る舞いが確認され、(ii) 共分散パラメータについては出現頻度が高い単語ほど広がりが大きくなることが確認された。

今後の研究課題としては、共分散パラメータの性質や、文や句を構成する単語の意味の変化の性質に関して、定量的に調査を進めていく予定である。

謝辞

本研究は、文部科学省科研費（23240018）、JST 戦略的創造研究推進事業「さきがけ」および、東北大学工学部 情報知能システム総合学科「Step-QI スクール」から部分的な支援を受けて行われた。

参考文献

[1] G. Dahl, R. Adams, and H. Larochelle. Training restricted boltzman machines on word observations. In *Proceedings of ICML*, 2012.

[2] K. Erk. Supporting inferences in semantic space: representing words as regions. In *Proceedings of IWCS*, 2009.

[3] N. T. Pham G. Dinu and M. Baroni. General estimation and evaluation of compositional distributional semantics models. In *the Workshop on Continuous Vector Space Models and their Compositionality*, 2013.

[4] G. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, Vol. 14, No. 8, pp. 1771–1800, 2002.

[5] J. Mitchell and M. Lapata. Vector-based models of semantic composition. In *ACL*, 2008.

[6] C. D. Manning R. Socher, B. huval and A. Y. Ng. Semantic compositionality through recursive matrix-vector spaces. In *EMNLP*, 2012.

[7] M. Geffet and I. Dagan The distributional inclusion hypotheses and lexical entailment. In *ACL*, 2005.